

# Bayesian Inference for the Multivariate Extended-Skew Normal Distribution

Mathieu Gerber\*

*Faculty of Business and Economics (HEC)*

*University of Lausanne, Switzerland*

*CREST*

Florian Pelgrin<sup>†</sup>

*EDHEC Business School, France*

*CIRANO*

The multivariate extended skew-normal distribution allows for accommodating raw data which are skewed and heavy tailed, and has at least three appealing statistical properties, namely closure under conditioning, affine transformations, and marginalization. In this paper we propose a Bayesian computational approach based on a sequential Monte Carlo (SMC) sampler to estimate such distributions. The practical implementation of each step of the algorithm is discussed and the elicitation of prior distributions takes into consideration some unusual behaviour of the likelihood function and the corresponding Fisher information matrix. Using Monte Carlo simulations, we provide strong evidence regarding the performances of the SMC sampler as well as some new insights regarding the parametrizations of the extended skew-normal distribution. A generalization to the extended skew-normal sample selection model is also presented. Finally we proceed with the analysis of two real datasets.

*Keywords:* Bayesian estimation; Bayes factor; Sequential Monte Carlo; Skew-elliptical distributions

---

\*Present address: Harvard University, Department of Statistics. Email: mathieugerber@fas.harvard.edu

<sup>†</sup>Email: florian.pelgrin@edhec.edu

# 1. Introduction

Recent years have seen a growing interest for flexible parametric families of multivariate distributions that can accommodate both the skewness and the kurtosis often observed on data. This is especially important, e.g., in health, finance and environmental data, which are often skewed and heavy tailed. For instance, these two features of health (-care) expenditures have fundamental implications in topics related to risk adjustments, program and treatment evaluations, or insurance choices (Manning et al., 2005). In this respect, the application of the skew-elliptical family of distributions (i.e., all non-symmetric distributions obtained from an elliptical distribution) has been put forward in the literature, and for good reasons. On the one hand, this family (or certain distributions of this family) has at least three appealing properties: closure under conditioning, affine transformations, and marginalization. On the other hand, these parametric distributions appear in the natural and important context of selection models (Heckman, 1976; Copas and Li, 1997; Arnold and Beaver, 2002). This last feature is particularly relevant in various research topics (e.g., economics, environmetrics or political sciences).

Within the class of skew-elliptical distributions, the extended skew-normal (ESN) distribution appears in different areas of statistical theory, e.g., Bayesian statistics (O'Hagan and Leonard, 1976), regression analysis (Copas and Li, 1997) or graphical models (Capitanio and Stanghellini, 2003). This generalization of the skew normal distribution, studied in the seminal paper of Azzalini (1985), has been developed and studied by Arnold et al. (1993); Arnold and Beaver (2000); Capitanio and Stanghellini (2003). Such a family of distributions may be obtained as a convolution between a multivariate normal random variable,  $\tilde{\mathbf{Y}}$ , and a truncated standard normal random variable,  $Z$ , say  $\mathbf{Y} \stackrel{d}{=} \tilde{\mathbf{Y}} + \mathbf{d}Z$ . The derivation and statistical properties of ESN distributions make it a natural candidate to model skewed and (leptokurtic, platikurtic) mesokurtic data generating processes. However, as pointed out by Arellano-Valle et al. (2006), statistical inference of skew-elliptical distributions, and thus of the extended skew-normal family of distributions, is still mostly unsolved (even in the univariate case).

Indeed, if Bayesian analysis of some families of skew-elliptical distributions has been proposed in the literature, they mainly focus on the skew-normal (or skew-t) distribution. In addition, Bayesian analysis of these distributions usually rely on objective prior-based methods, Gibbs sampling or population Monte Carlo. In particular, Liseo and Loperfido

(2006) consider a Bayesian estimation of the univariate skew-normal distribution based on objective priors whereas Wiper et al. (2008) analyse the half-normal and half-t cases, and Branco et al. (2013) focus on the skew-t distribution. Cabral et al. (2012) propose a full Bayesian estimation of a mixture of skew-normal densities while Fr  wirth-Schnatter and Pyne (2010) provide a Gibbs sampler to estimate a mixture of skew-normal and skew-t densities. As an alternative to the Gibbs sampler, Liseo and Parisi (2013) advocate the application of a Population Monte Carlo (PMC) algorithm for missing data (Celeux et al., 2004) in order to sample from the posterior distribution of the skew-normal model.

In this paper we propose a Bayesian computational method to estimate the extended (multivariate) skew-normal distribution. The Bayesian approach for this family of distributions is motivated by some severe anomalies of the likelihood function and some identification issues. Notably, and as shown in this paper, the maximum likelihood estimator may not be uniquely defined for univariate ESN distributions. In a Bayesian approach, these anomalies can be tackled to some extent by using a suitable elicitation of prior distributions. In the light of the properties of the likelihood function, we make use of a (tempered) sequential Monte Carlo sampler (Del Moral et al., 2006) rather than a Monte Carlo Markov chain (MCMC) algorithm. Briefly speaking, sequential Monte Carlo (SMC) iterates importance sampling steps, resampling steps and Markov kernel transitions in order to recursively approximate a sequence of distributions by making use of a sequence of weighted particle systems. Relative to MCMC algorithms, SMC might be called for at least three arguments. On the one hand, the great generality of SMC allows to build efficient algorithms in order to sample from complicated distributions and thus to overcome some distortions of the likelihood functions of the skew-normal distribution. On the other hand, compared to MCMC methods, it is easier to make SMC algorithms adaptive in the sense that they can be adjusted sequentially and automatically to the problem at hand, and the evidence or marginal likelihood of data can be derived formally. Finally, due to the convolution representation of the ESN distribution presented above, a natural idea to sample from the posterior distribution would be to implement a Gibbs sampler in which the hidden random variable  $Z$  is an extra parameter. However, in the case of ESN models, the support of the hidden variable  $Z$  depends on the parameters of interest and thus the posterior distribution in the augmented space does not satisfy the positivity condition (see Robert and Casella, 2004, chapter 9).

The rest of the paper is organized as follows. Section 2 defines two equivalent rep-

representations of extended skew-normal random vectors, review some useful properties of this class of distributions, and discuss some unpleasant features of the maximum likelihood function. Section 3 proposes some prior distributions which take into consideration these anomalies of the likelihood function and describes the proposed tempered sequential Monte Carlo algorithm. Section 4 presents some Monte Carlo simulations regarding the inference of univariate ESN distributions and of some regressions with missing data. Moreover we discuss the testing and model selection problems. Section 5 deals with two applications, namely the distribution of transfer fees of soccer players in major European leagues and the bivariate distribution of two financial returns (Liseo and Parisi, 2013). The last section provides some concluding remarks.

## 2. The extended skew-normal distribution

In this section we first define the extended skew-normal (ESN) distribution using two different parametrizations. Then, we review some appealing properties of this class of distributions, especially in the light of the subsequent derivations of this paper. Finally, we provide a new theoretical justification for the unsatisfactory behaviour of the maximum likelihood estimator of the ESN distribution.

### 2.1. Definition and main properties

We consider two parametrizations of the ESN distribution. The first parametrization, denoted P1, is based on hidden truncation (and/or selective reporting) using normal component densities whereas the second parametrization, denoted P2, rests on the convolution of a multivariate normal distribution with a truncated standard normal variable.

**Definition 1.** *A random vector  $\mathbf{Y}$  is said to have a  $d$ -dimensional extended skew-normal distribution, denoted  $\mathbf{Y} \sim \mathcal{ESN}_d^{(P1)}(\boldsymbol{\xi}, \Sigma, \boldsymbol{\alpha}, \lambda)$ , with covariance (correlation) matrix  $\Sigma$ , shape parameter  $\boldsymbol{\alpha}$ , and shift parameter  $\lambda$ , if*

$$\mathbf{Y} \stackrel{d}{=} (\boldsymbol{\xi} + \tilde{\mathbf{Y}}_1 | \lambda + \boldsymbol{\alpha}' \tilde{\mathbf{Y}}_1 > Z_1), \quad \begin{pmatrix} \tilde{\mathbf{Y}}_1 \\ Z_1 \end{pmatrix} \sim \mathcal{N}_{d+1} \left( \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \right) \quad (1)$$

where  $\mathcal{N}_{d+1}(\boldsymbol{\mu}, B)$  denotes the  $(d+1)$ -dimensional Gaussian distribution with mean  $\boldsymbol{\mu}$  and variance-covariance matrix  $B$ . Its density function is defined to be:

$$f_Y(\mathbf{y}) = \phi_d(\mathbf{y}, \boldsymbol{\xi}, \Sigma) \frac{\Phi(\lambda + \boldsymbol{\alpha}'(\mathbf{y} - \boldsymbol{\xi}))}{\Phi(\lambda/c_0)}, \quad c_0 = \sqrt{1 + \boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}} \quad (2)$$

where  $\phi_d(\cdot, \boldsymbol{\mu}, B)$  is the density of the  $d$ -dimensional normal distribution with mean  $\boldsymbol{\mu}$  and covariance (correlation) matrix  $B$  and  $\Phi(\cdot)$  is the cumulative density function (cdf) of the  $\mathcal{N}_1(0, 1)$  distribution.

On the other hand, the ESN distribution can be defined from a convolution.

**Definition 2.** A random vector  $\mathbf{Y}$  is said to have a  $d$ -dimensional extended skew-normal distribution, denoted  $\mathbf{Y} \sim \mathcal{ESN}_d^{(P2)}(\boldsymbol{\xi}, \Omega, \mathbf{d}, c)$ , if

$$\mathbf{Y} \stackrel{d}{=} \tilde{\mathbf{Y}}_3 + \mathbf{d}Z_3$$

where  $-Z_3 \sim \mathcal{TN}_c(0, 1)$ , the  $\mathcal{N}_1(0, 1)$  distribution truncated to  $(-\infty, c]$ , and  $\tilde{\mathbf{Y}}_3 \sim \mathcal{N}_d(\mathbf{0}, \Omega)$ . Its probability density function is defined to be:

$$f_Y(\mathbf{y}) = \phi_d(\mathbf{y}, \boldsymbol{\xi}, \Omega + \mathbf{d}\mathbf{d}') \frac{\Phi\left(c_0 \left\{c + \mathbf{d}'[\Omega + \mathbf{d}\mathbf{d}']^{-1}(\mathbf{y} - \boldsymbol{\xi})\right\}\right)}{\Phi(c)}.$$

Several points are worth commenting. First, the ESN distribution belongs to the families of skew-elliptical distributions proposed by Arnold and Beaver (2002), Domínguez-Molina et al. (2003), Fang (2003), and Arellano-Valle and Genton (2010). Alternatively, using P2, the ESN distribution belongs to the family of distributions proposed by Sahu et al. (2003). Irrespective of the parametrization, the ESN distribution generalizes the multivariate skew-normal distribution (Azzalini and Dalla Valle, 1996) and thus the Gaussian distribution. More specifically, when the shift parameter  $\lambda$  is set to zero, one obtains the (multivariate) SN distribution. On the other hand, the standard normal distribution results from the nullity of the shape parameter vector  $\boldsymbol{\alpha}$ . As explained in Section 2.2, this constraint on the shape parameter vector has some key implications on inference. Indeed, the Fisher information matrix of the ESN (and of the SN) distribution is singular, preventing a straightforward application of standard likelihood-based methods to test the null hypothesis of normality. The problem is even made worse by the parameter  $\lambda$ , which indexes the distribution in the case of non-normality (nuisance parameter).

Second, the choice of the parametrization might be critical for the estimation of ESN distributions since, in a Bayesian perspective, different parametrizations lead to alternative choices of prior distributions and thus different models (see Section 3). Third, one key feature of the ESN distribution over the SN distribution is that the former has an extra parameter that allows for a larger range of values for skewness and kurtosis and

thus for more flexibility to accommodate real data. For instance, using the moment generating function of Domínguez-Molina et al. (2003), one can provide evidence with a numerical analysis of the univariate ESN that the skewness coefficient is bounded by 2 (in absolute value) while the kurtosis coefficient varies roughly between 2.75 and 7. In contrast, Azzalini (1985) points out that the skewness is smaller (in absolute value) than 0.995 and that the kurtosis lies between 3 and 3.87 in the case of the skew-normal distribution.

Fourth, the ESN distribution has three familiar and useful properties, especially for regression-type models. It is closed under affine transformations, conditioning and marginalization. On the one hand, ESN random vectors share the affine transformation of normal random vectors. In particular, let  $A$  denote an  $d \times d$  non-singular matrix and  $\tilde{\xi} \in \mathbb{R}^d$ . Then, taking (1), one obtains  $\tilde{\xi} + A'Y \sim \mathcal{ESN}_d^{(P1)}(\tilde{\xi} + A'\xi, A'\Sigma A, A^{-1}\alpha, \lambda)$ . On the other hand, if an ESN vector is partitioned into two components, the conditional distribution of one component given the other is extended skew-normal and each component is marginally extended skew-normal. For sake of completeness, Proposition 1 due to Fang (2003) and Domínguez-Molina et al. (2003) reports the closure of the ESN distribution under conditioning and marginalization.

**Proposition 1.** Assume that  $Y \sim \mathcal{ESN}_d^{(P1)}(\xi, \Sigma, \alpha, \lambda)$ . Partition  $Y$ ,  $\xi$ ,  $\alpha$  and  $\Sigma$  as  $Y = (Y_1, Y_2)'$ ,  $\epsilon = (\epsilon_1, \epsilon_2)'$ ,  $\alpha = (\alpha_1, \alpha_2)'$  and  $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$  where  $Y_i$ ,  $\xi_i$  and  $\alpha_i$  are  $m_i \times 1$  and  $\Sigma_{ii}$  is  $m_i \times m_i$ . Then,

$$Y_i \sim \mathcal{ESN}_{m_i}^{(P1)}(\xi_i, \Sigma_{ii}, c_i \tilde{\alpha}_i, c_i \lambda), \quad (Y_i | Y_j = y_j) \sim \mathcal{ESN}_{m_i}^{(P1)}(\xi_i^c, \Sigma_{ii.1}, \alpha_i, \lambda_i)$$

where  $c_i = (1 + \alpha_j' \Sigma_{i.1} \alpha_j)^{-1/2}$ ,  $\xi_i^c = \xi_i + \Sigma_{ij} \Sigma_{jj}^{-1} (y_j - \xi_j)$ ,  $\Sigma_{ii.1} = \Sigma_{ii} - \Sigma_{ij} \Sigma_{jj}^{-1} \Sigma_{ji}$ ,  $\tilde{\alpha}_i = \alpha_i + \Sigma_{ii}^{-1} \Sigma_{ij} \alpha_j$  and  $\lambda_i = \lambda + \tilde{\alpha}_j (y_j - \xi_j)$ .

Finally, the stochastic representation (1) of ESN random vectors leads to the following expression of the cumulative density function (henceforth, cdf):

$$\mathbb{P}(Y \leq y) = \frac{\Phi_{d+1}(y - \xi, \Sigma, \alpha, \lambda)}{\Phi(\lambda/c_0)}$$

with  $\Phi_{d+1}(\mathbf{a}, \Sigma, \alpha, \lambda) = \mathbb{P}(\tilde{Y}_2 \leq \mathbf{a}, Z_2 \leq \lambda)$  and where

$$(\tilde{Y}_2, Z_2) \sim \mathcal{N}_{d+1} \left( \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & -\Sigma \alpha' \\ -\alpha' \Sigma & c_0^2 \end{pmatrix} \right).$$

Notably, the evaluation of the cdf of the  $d$ -dimensional ESN distribution has the same complexity as the computation of the cdf of the  $(d+1)$ -dimensional Gaussian distribution, for which efficient methods are available (e.g., see Huguenin et al., 2014). It turns to be very useful in practice since, for instance, the cdf of the ESN distribution arises naturally when deriving the expression of the likelihood function in the presence of missing data (see Section 4.2).

## 2.2. Log-likelihood function

Since our methodology rests on Bayesian estimation and thus on the posterior distribution associated to the ESN-based model, it is fundamental to study the statistical properties of the likelihood function. This might provide some useful insights in order to determine the prior distribution and thus challenge some identified anomalies regarding the likelihood function (i.e., to correct at least partially the odd behaviour of the likelihood function with external information). For sake of exposition, we concentrate on the univariate ESN distribution.

Maximum likelihood estimation of ESN distributions is challenging and quite difficult to manage. More specifically, it is widely acknowledged that (i) there are no closed form expressions for the maximum likelihood estimator (MLE), (ii) the MLE of  $\alpha$  can be infinite even in very simple settings, (iii) the multimodality of the log-likelihood profile (and thus local solutions) can not be ruled out and (iv) there exists an inflexion point at  $\alpha = 0$ . In particular, the Fisher information matrix tends to be singular as  $\alpha$  goes toward zero irrespective of the  $\lambda$  parameter. Note that, in this case, ESN distributions are no longer indexed by the normal cumulative density functions and, consequently, the rank of the information matrix might be at least two less than its full rank. On the other hand, the presence of a stationary point (e.g., using the profile log-likelihood for the  $\alpha$  parameter) and of multiple modes generally cause numerical issues.

While these issues have been outlined in the literature, to the best of our knowledge, there is not yet a formal proof of the near unidentifiability of the log-likelihood function and the  $\lambda$  parameter. Therefore, we show in Proposition 2 that the presence of the shift parameter  $\lambda$  in P1 might lead to local maxima for the maximum likelihood estimator of the univariate extended skew-normal distribution. Indeed, irrespective of the data and for all  $l \in \mathbb{R}$ , the ESN distribution admits a stationary point at  $\theta_{n,G}^l := (\boldsymbol{\xi}_{n,G}, \Sigma_{n,G}, \mathbf{0}_d, l)$ , where  $\boldsymbol{\xi}_{n,G}$  and  $\Sigma_{n,G}$  are the MLE of  $\boldsymbol{\xi}$  and  $\Sigma$  under the Gaussian assumption. In so

doing, if this stationary point is an inflexion point when we impose the  $\lambda$  parameter to be zero (Azzalini and Capitanio, 1998), the problem becomes even more severe when  $\lambda$  is a free parameter as stated in Proposition 2.

**Proposition 2.** *Let  $Y_1, \dots, Y_n$  be  $n$  i.i.d. random variables,  $Y_1 \sim \mathcal{ESN}_1^{(P1)}(\xi, \sigma^2, \alpha, \lambda)$  with  $\alpha \neq 0$ . Let  $\theta_{n,G}^l = (\xi_{n,G}, \sigma_{n,G}^2, 0, l)$  with  $\xi_{n,G} = \frac{1}{n} \sum_{i=1}^n Y_i$  and  $\sigma_{n,G}^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 - \xi_{n,G}^2$ . Let  $L_n(\theta)$  denote the log-likelihood function. Then,*

1. *With probability one, there exists a  $l^* \in \mathbb{R}$  such that  $L_n(\theta_{n,G}^l)$  is a local maximum of  $L_n(\cdot)$  for all  $l \leq l^*$ ;*
2. *With strictly positive probability,  $L_n(\theta_{n,G}^l) = L_n(\theta_n)$ ,  $l \in \mathbb{R}$ , where  $\theta_n \neq \theta_{n,G}^l$  is a global maximizer of  $L_n(\cdot)$ .*

See Appendix A for a proof.

The first result of Proposition 2 has an intuitive interpretation. When  $\alpha = 0$ , the value of the log-likelihood function is insensitive to any change of the  $\lambda$  parameter and thus any small deviation of  $\alpha$  leads to large deviations from the true log-likelihood value (since the estimate  $\lambda$  was initially far from its true unknown value). Consequently, a small deviation from  $\theta_{n,G}^l$  in any direction reduces the value of the likelihood. The second part of Proposition 2 is more puzzling because it implies that, with a positive (but decreasing with  $n$ ) probability, the likelihood function does not allow to discriminate between the Gaussian and the ESN model. This is a particularly severe anomaly of the likelihood function because it implies that the MLE might be not uniquely defined.

### 3. Bayesian analysis of the ESN distribution

In this section we first discuss the elicitation of prior distributions and then explain how to estimate the parameters of the two parameterizations of ESN distributions using Sequential Monte Carlo (Del Moral et al., 2006).

#### 3.1. A default Prior specification

In contrast to the standard approach of default prior distributions, and in the spirit of Gelman et al. (2008), we propose a prior specification that embeds enough information to circumvent the anomalies of the log-likelihood function listed in Section 2.2.



On the one hand, the  $(\boldsymbol{\xi}, \Sigma, \boldsymbol{\alpha}, \lambda)$ -parametrization (P1) of ESN random vectors must tackle two issues, namely the potential existence of multiple modes and the identification (estimation) of the truncation point, that are related to the identification of the  $\lambda$  parameter. First, the multi-modality of the log-likelihood function might be attenuated by setting a prior that assigns less weight on very negative values of  $\lambda$ . Second, as argued in Section 2.2, values of  $\lambda$  such that the truncation point exceeds a certain threshold, say  $|\lambda|/c_0 > 2$ , are difficult to identify and therefore, both to avoid extreme estimates of  $\lambda$  and to facilitate its identification in this region of the parameter space, it is important to choose a prior  $\pi(d\lambda|\Sigma, \boldsymbol{\alpha})$  that puts small weights on  $\{l \in \mathbb{R} : |l|/c_0 > 2\}$ . In so doing, we propose to consider a conditional normal prior distribution with mean zero and variance  $c_0^2$ , i.e.  $\lambda|(\Sigma, \boldsymbol{\alpha}) \sim \mathcal{N}_1(0, c_0^2)$ . This naturally leads to a normal-inverse Wishart distribution as a prior for  $(\boldsymbol{\xi}, \Sigma)$ , which is the conjugate prior for Gaussian models (e.g., see Gelman et al., 2004). Note that it turns to ease the Bayesian model selection procedure (see Section 3.2.3). Hence,

$$\pi(\boldsymbol{\xi}, \Sigma|\boldsymbol{\alpha}) \propto \exp\left(-\frac{1}{2}\text{tr}(V\Sigma^{-1}) - \frac{\kappa}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}_0)' \Sigma^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}_0)\right) |\Sigma|^{-\frac{\nu+d+2}{2}} \quad (3)$$

where  $V$  is a  $d \times d$  positive definite matrix,  $\kappa$  and  $\nu$  are real such that  $\nu > d + 3$ . This last condition ensures that the mean of the prior distribution of  $\Sigma$  exists and that all its components has finite variance. Finally, one can choose a vague prior for  $\boldsymbol{\alpha}$ , e.g.  $\boldsymbol{\alpha} \sim \mathcal{N}_d(\boldsymbol{\mu}_\alpha, \sigma_\alpha^2 I_d)$  with  $\sigma_\alpha^2$  large and  $I_d$  the  $d \times d$  identity matrix. In practice, it is likely to have information on the sign of  $\alpha_i$ ,  $i = 1, \dots, d$ , through information about the asymmetry of the full conditional distribution of  $Y_i$  (see Proposition 1). This prior knowledge can be incorporated in the Bayesian analysis by taking  $\boldsymbol{\mu}_\alpha \neq \mathbf{0}_d$ .

On the other hand, the  $(\boldsymbol{\xi}, \Omega, \mathbf{d}, c)$ -parametrization shares the same issues as the P1-parametrization since  $c = \lambda/c_0$ . In addition, since an ESN random vector  $\mathbf{Y}$  is defined by  $\mathbf{Y} \stackrel{d}{=} \boldsymbol{\xi} + \mathbf{d}Z_3 + \Omega^{1/2}\tilde{\mathbf{Y}}_3$ , where  $-Z_3 \sim \mathcal{TN}_c(0, 1)$  and  $\tilde{\mathbf{Y}}_3 \sim \mathcal{N}_d(\mathbf{0}_d, I_d)$ , the convolution representation of ESN random vectors leads to an additional identification problem that arises when  $\Omega$  is “large” or “small” relative to  $\mathbf{d}$ —more variability of one of the convolution-based density is obtained at the expense of weak identification of the other. In this respect, we assume that  $\mathbf{d}|(\boldsymbol{\xi}, \Omega) \sim \mathcal{N}_d(\boldsymbol{\mu}_\mathbf{d}, \kappa_\mathbf{d}^{-1}\Omega)$ , with  $\kappa_\mathbf{d} = 2(\sigma_\alpha^2(\tilde{\nu} - d - 1))^{-1}$ , yielding, on average, the same variance for both  $\mathbf{d}$  and  $\boldsymbol{\alpha}$ . The choice of a Gaussian distribution for  $(\mathbf{d}|\Omega)$  is motivated by the fact that, together with the assumption that the prior distribution of  $(\boldsymbol{\xi}, \Omega)$  is the normal-inverse Wishart distribution

$\pi(\boldsymbol{\xi}, \Sigma | \tilde{\boldsymbol{\xi}}_0, \tilde{\kappa}, \tilde{\nu}, \tilde{V})$ , we can easily implement a Gibbs sampler when  $c$  is known (Sahu et al., 2003). Say differently, the Gaussian prior for  $\mathbf{d}$  and the normal-inverse Wishart prior for  $(\boldsymbol{\xi}, \Omega)$  are some natural candidates for the SN distribution of Sahu et al. (2003). Since  $\Sigma - \Omega = \Sigma \boldsymbol{\alpha} \boldsymbol{\alpha}' \Sigma / c_0^2$ , which is a positive definite matrix, we choose  $(\tilde{\nu}, \tilde{V})$  such that the inverse Wishart distribution  $\mathcal{W}^{-1}(\tilde{V}, \tilde{\nu})$  gives more weight to “small” values than the  $\mathcal{W}^{-1}(V, \nu)$  distribution. This can be done by taking  $\tilde{V}$  such that  $V - \tilde{V}$  is positive definite and  $\tilde{\nu} \geq \nu$ . In this case, the difference between the mode under  $(\tilde{\nu}, \tilde{V})$  and the mode under  $(\nu, V)$  is negative definite. This also holds for the mean.

## 3.2. A SMC sampler for multivariate ESN distributions

### 3.2.1. General description

Let  $\theta$  be the vector whose components are the parameters of the model (either under P1 or under P2),  $f(\mathbf{z}_{1:n} | \theta)$  be the likelihood function, where  $\mathbf{z}_{1:n} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$  is the set of observations, and  $\pi(\theta)$  be the prior distribution of the parameters, which is either  $\pi_{P1}(\theta)$  under P1 or  $\pi_{P2}(\theta)$  under P2. Using these notations, the posterior distribution we want to sample from is given by:

$$\pi(\theta | \mathbf{z}_{1:n}) \propto f(\mathbf{z}_{1:n} | \theta) \pi(\theta).$$

As pointed out by Del Moral et al. (2006), sequential Monte Carlo samplers are relevant when there is no fully-eligible proposal distribution, say  $\eta_1(\theta)$ , in order to implement the importance sampler. The SMC sampler requires to define a sequence of distributions  $\{\pi_t(\theta)\}_{t=1}^T$  such that (i)  $\pi_T(\theta) = \pi(\theta | \mathbf{z}_{1:n})$  and (ii)  $\pi_1(\theta) = \eta_1(\theta)$ , with  $\eta_1$  a distribution we can easily sample from. This sequence of intermediary distributions is purely instrumental and could be defined by making use of an appropriate real sequence of so-called temperatures  $\{\rho_t\}_{t=1}^T$ , increasing from zero to one. Following Gelman and Meng (1998), Neal (2001) and Del Moral et al. (2006), we consider the geometric bridge:

$$\pi_t(\theta) \propto \eta_1(\theta)^{1-\rho_t} \pi(\theta | \mathbf{z}_{1:n})^{\rho_t}.$$

The basic idea of a SMC algorithm is first to sample  $N \geq 1$  particles  $\theta_1^m$  from the initial distribution  $\pi_1$  in order to obtain a Monte Carlo approximation  $\pi_1^N = N^{-1} \sum_{m=1}^N \delta_{\theta_1^m}$  of  $\pi_1$ . Then, using resampling and propagation steps, SMC uses the approximation  $\pi_1^N$  of  $\pi_1$  to construct  $\pi_2^N = \sum_{m=1}^N W_2^m \delta_{\theta_2^m}$ , a Monte Carlo approximation of  $\pi_2$ . Informally, if

$\pi_1^N$  is a good approximation of  $\pi_1$  and if  $\pi_2$  is close to  $\pi_1$ , then one may expect  $\pi_2^N$  to be close to  $\pi_2$ , and so on.

More precisely, suppose that, for a  $t \in \{2, \dots, T\}$ , one has at hand a sample  $\{\theta_t^m\}_{m=1}^N$  such that:

$$\frac{1}{N} \sum_{m=1}^N \delta_{\theta_t^m}(\mathrm{d}\theta) \approx \pi_t(\theta) \mathrm{d}\theta.$$

Then, we can approximate  $\pi_{t+1}$  by the empirical distribution

$$\sum_{m=1}^N W_{t+1}^m(\rho_{t+1}) \delta_{\theta_t^m}(\mathrm{d}\theta)$$

where the corresponding importance functions  $W_{t+1}^m(\cdot)$  are defined to be:

$$W_{t+1}^m(\rho) = \frac{w_{t+1}(\theta_t^m, \rho)}{\sum_{j=1}^N w_{t+1}(\theta_t^j, \rho)}, \quad w_{t+1}(\theta, \rho) = \left[ \frac{\pi(\theta | \mathbf{z}_{1:n})}{\eta_1(\theta)} \right]^{\rho - \rho_t}.$$

Note that  $\rho_{t+1} - \rho_t$  measures the step length at time  $t + 1$  so that, the larger the difference, the more the accuracy of the importance weighting worsens. To control such a degeneracy, we consider a procedure to determine a suitable sequence of  $\{\rho_t\}_{t=1}^T$  through the effective sample size criterion. More specifically, instead of regarding  $T$  and the set  $\{\rho_t\}_{t=1}^T$  as parameters of the algorithm, we view them as self-tuning parameters using the method proposed by Schäfer and Chopin (2013). Given a value of  $\rho_t$  and a sample  $\{\theta_t^m\}_{m=1}^N$  that approximates  $\pi_t$ , we compute the largest value of  $\rho \in (\rho_t, 1]$  such that the particle system  $\{\theta_t^m\}_{m=1}^N$ , once being properly weighted, allows to approximate “reasonably well” the probability distribution  $\pi_\rho \propto \eta_1^{1-\rho} \pi^\rho$  through the effective sample size criterion (Liu and Chen, 1995):

$$\text{ESS}_t(\rho) = \left[ \sum_{m=1}^N W_t^m(\rho)^2 \right]^{-1}$$

where, by definition,  $W_t^m(\rho)$  is the weight assigned to  $\theta_t^m$  to target  $\pi_\rho$ . If the effective sample size equals  $N$ , the interpretation is that the weights are equally balanced and that all  $N$  particles are equally contributing to the estimation. Then,  $\rho_{t+1}$  is defined as the minimum between 1 and  $\rho_{t+1}^*$  with:

$$\rho_{t+1}^* = \sup \{ \rho > \rho_t : \text{ESS}_t(\rho) \geq \beta \}$$

where  $\beta$  is a pre-specified threshold, say  $\beta = N/2$ . The fixed value  $\rho_{t+1}^*$  can be obtained by solving the equation  $\text{ESS}_t(\rho) = \beta$  using the bi-sectional search algorithm of Schäfer and Chopin (2013) (see Algorithm 2).

---

**Algorithm 1** Tempering Sequential Monte Carlo Sampler

---

Operations must be performed for all  $m = 1, \dots, N$ .

Initialization

Set  $t = 2$  and  $\rho_1 = 0$ .

Generate  $\theta_1^m \sim \eta_1(d\theta)$  and compute  $W^m(\rho_1)$ .

**while**  $\rho_{t-1} < 1$  **do**

    Compute  $\rho_t$  using Algorithm 2 with inputs  $\rho_{t-1}$  and  $\{\theta_{t-1}^m\}_{m=1}^N$ .

Resampling: Generate  $a_{t-1}^m = F_{t,N}^{-1}(u_t^m)$  where  $u_t^m = \frac{m-1+u_t}{N}$ ,  $u_t \sim \mathcal{U}((0, 1))$  and

$$F_{t,N}(i) = \sum_{m=1}^N W_t^m(\rho_t) \mathbb{I}(m \leq i).$$

Propagation: Generate  $\theta_t^m \sim K_t^N(\theta_{t-1}^{a_{t-1}^m}, d\theta)$ .

    Set  $t \leftarrow t + 1$ .

**end while**

---

At every  $\rho_t$ , a resampling step, using the systematic resampling method of Carpenter et al. (1999), is first performed in order to suppress particles that are in the region of the parameter space that receives very little mass from  $\pi_t$ . Say differently, the particles with the largest weights have multiplied whereas those with the smallest weights have vanished after the resampling step. Then, to restore particle diversity, new particles are generated from a Markov Kernel  $K_t^N(\theta', d\theta)$  with invariant distribution  $\pi_t$  (see further).

The complete procedure is summarized in Algorithm 1. Any operation involving the superscript  $m$  (respectively, subscript  $t$ ) must be understood as performed for  $m \in 1 : N$  (respectively,  $t \in 0 : T$ ) where  $N$  (respectively,  $T$ ) is the total number of particles (respectively, number of iterations). Note that  $n$  denotes the sample size. In addition, the procedure to find the step length is described in Algorithm 2.

### 3.2.2. Implementation

In our implementation we follow the usual approach and take for  $K_t^N(\theta', d\theta)$  the Markov kernel that corresponds to  $\tau$  steps of the Gaussian random-walk Metropolis-Hastings algorithm with variance-covariance matrix given by  $c_s \Omega_t^N$ . The constant  $c_s > 0$  is a scale factor such that the acceptance rate of the kernel lies in the range  $[0.2, 0.6]$  while  $\Omega_t^N$  is a particle-based estimation of the variance-covariance matrix that corresponds to the distribution  $\pi_t$ .

The initial distribution  $\eta_1$  is another critical element for the speed of convergence

---

**Algorithm 2** Find step length using Schäfer and Chopin (2013)

---

**Input:**  $\epsilon, \rho, \{\theta^m\}_{m=1}^N$ .  
 $l \leftarrow 0, u \leftarrow 1.05, \delta \leftarrow 0.05$ .  
**while**  $|u - l| \geq \epsilon$  and  $l \leq 1 - \rho$  **do**  
    **if**  $\left[\sum_{m=1}^N W^m(\rho + \delta)^2\right]^{-1} < N/2$  **then**  
         $u \leftarrow \delta, \delta \leftarrow (\delta + l)/2$   
    **else**  
         $l \leftarrow \delta, \delta \leftarrow (\delta + u)/2$   
    **end if**  
**end while**  
**Return**  $\min(\rho + a, 1)$ .

---

of the algorithm and for the precision of the estimates. The first obvious option is to take the prior distribution, so that the SMC sampler moves simulations from the prior to simulations from the posterior distribution. Nevertheless, starting the SMC sampler with simulations from the prior can lead to a very low convergence rate of the algorithm and some large Monte-Carlo errors since there is no reason for the prior to be close to the posterior distribution. A better approach consists in initializing the sampler with an approximation of the target distribution from which we can easily sample. When one can maximize the posterior distribution, this is effectively done by a Laplace approximation. In this case,  $\eta_1$  would be a normal distribution with mean  $\mathbf{m}_1$  and covariance matrix  $\Sigma_1$ , where  $\mathbf{m}_1$  is set to the posterior mode and  $\Sigma_1$  is equal to minus the inverse of the Hessian matrix evaluated at the posterior mode. In some settings (see Section 4), the numerical maximization of the posterior distribution might be particularly troublesome. In this case, we use a pilot run of a Gaussian random walk Metropolis-Hastings algorithm to get an estimate  $\hat{\mathbf{m}}$  of the posterior mean and an estimate  $\hat{\Sigma}$  of the posterior covariance matrix, and we define  $\eta_1$  as the a normal distribution with mean  $\hat{\mathbf{m}}$  and covariance matrix  $\hat{\Sigma}$ .

### 3.2.3. Discussion

Using Algorithm 1, one can obtain estimates of the target distributions and the normalizing constants directly from the variables generated by the sampler. Indeed, at the end of iteration  $T$ , an approximation of the target distribution  $\pi(\theta|\mathbf{z}_{1:n})$  is given by:

$$\pi_T^N(d\theta) = \frac{1}{N} \sum_{m=1}^N \delta_{\theta_T^m}(d\theta).$$

Moreover an estimate of the normalizing constant  $Z_T$  of the posterior distribution  $\pi_T(\theta)$  can be obtained as follows. Let  $Z_t$  be the normalizing constant of  $\pi_t$ . Then, we can estimate of  $Z_T/Z_1$  by (Del Moral et al., 2006):

$$\frac{\widehat{Z_T}}{\widehat{Z_1}} = \prod_{t=2}^T \frac{\widehat{Z_t}}{\widehat{Z_{t-1}}}, \quad \frac{\widehat{Z_t}}{\widehat{Z_{t-1}}} = \sum_{m=1}^N W_{t-1}^m(\rho_{t-1}) \left[ \frac{\pi(\theta_{t-1}^m | \mathbf{z}_{1:n})}{\eta_1(\theta_{t-1}^m)} \right]^{\rho_t - \rho_{t-1}}, \quad t \geq 2.$$

A question of particular interest is whether the SN or the Gaussian distributions are more appropriate than the ESN distribution. In a Bayesian framework, the answer to this question is obtained by comparing the evidence, or marginal likelihood of the data, between the competing models. More specifically, consider the general test  $H_0 : \theta = \theta^0$  against  $H_1 : \theta \neq \theta^0$ , where  $\theta^0$  is the vector of parameters under the null hypothesis. In this respect, we make use of the Bayes factor defined by:

$$B_{10} = \frac{m_1(\mathbf{z}_{1:n})}{m_0(\mathbf{z}_{1:n})}$$

where  $\mathbf{z}_{1:n}$  is the observations and where  $m_i(\mathbf{z}_{1:n}) = \int f_i(\mathbf{z}_{1:n} | \theta) \pi_i(d\theta)$  is the evidence of model  $i \in \{0, 1\}$ , with  $f_i(\mathbf{z}_{1:n} | \theta)$  and  $\pi_i(d\theta)$  the corresponding likelihood and the prior distribution.

It is well known (Morin et al., 2013) that, if the competing models  $i \in \{0, 1\}$  are regular, then the Bayes factor is a consistent criterion to discriminate between  $H_1$  and  $H_0$ . However, and as discussed in Section 2.2, if we wrongly assume that data are generated by some ESN distributions when the true underlying model is Gaussian, then the Fisher information matrix is singular and therefore there is no theoretical guarantee that the Bayes factor selects asymptotically the true model. We leave this issue for further research and rather assess the Bayes factor reliability through Monte Carlo simulations in Section 4.

At this stage it is worth noting that testing the ESN distribution against the Gaussian model is straightforward. Indeed, the evidence under the ESN distribution can be directly obtained as a by-product of Algorithm 1, as explained above, while that under the Gaussian distribution can be computed explicitly thanks to the Gaussian conjugate prior (3) for  $\xi$  and  $\Sigma$  (see e.g. Gelman et al., 2004):

$$m_0(\mathbf{z}_{1:n}) = \frac{1}{\pi^{nd/2}} \frac{\Gamma_d(\nu_n/2)}{\Gamma_d(\nu/2)} \frac{|V|^{\nu/2}}{|V_n|^{\nu_n/2}} \left( \frac{\kappa}{\kappa_n} \right)^{d/2}$$

where

$$\kappa_n = \kappa + n, \quad \nu_n = \nu + n, \quad V_n = V + \sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}}_n)(\mathbf{z}_i - \bar{\mathbf{z}}_n)^t + \frac{\kappa n}{\kappa + n} (\bar{\mathbf{z}}_n - \boldsymbol{\xi}^0)(\bar{\mathbf{z}}_n - \boldsymbol{\xi}^0)'. \quad (3)$$

Finally, note that one key feature of SMC algorithms is their flexibility. Indeed, the implementation of Algorithm 1 only requires to be able to evaluate the likelihood function. The Bayesian methodology developed in this section can therefore be easily modified to carry out parameter inference in (complicated) parametric models based on the ESN distribution. This point is illustrated in Section 4.2 where we apply the proposed methodology on an ESN sample selection model.

## 4. Numerical study

In this section we provide some Monte Carlo simulations in order to assess the performances of the proposed Bayesian approach and the behaviour of the posterior distribution. We consider two main data generating processes: (1) IID univariate extended skew-normal random variables, and (2) an extended skew-normal sample selection model (ESNSM). For the IID setting, the SMC sampler is initialized with a Laplace approximation of the posterior distribution while, for the ESNSM, the maximization of the posterior distribution turns out to be too sensitive to the choice of initial values in order to be useful in the construction of a good approximation of the posterior distribution. In that case, and as described above, we calibrate the initial distribution of the sampler using 10 000 iterations of a pilot Metropolis-Hastings algorithm. Finally, in all of the simulations presented below, the propagation step of the tempered sequential Monte Carlo algorithm is based on  $\tau = 3$  iterations of the Gaussian random walk Metropolis-Hastings kernel described in Section 3.2.

### 4.1. Example 1: IID univariate ESN random variables

We first consider a sample of IID ESN random variables. To study the implications of the parametrization of ESN distributions, we use two data generating processes:

$$Z_1, \dots, Z_n \sim \mathcal{ESN}_1^{(P1)}(2, 6, 5, -2) \quad (4)$$

and

$$Z_1, \dots, Z_n \sim \mathcal{ESN}_1^{(P2)}(2, 1, 5, -0.8) \quad (5)$$

where the sample size  $n$  is successively 1 000, 5 000, and 10 000. The variance, skewness and kurtosis of the first ESN distribution (4) are respectively given by 2, 1, 4 whereas those of the second ESN distribution (5) are respectively given by 6.60, 0.99, and 4.28. For all of the simulations, the parameters for the prior distributions, defined in Section 3.1, are set as follows:  $\kappa = 0.1$ ,  $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = \boldsymbol{\mu}_{\mathbf{d}} = \boldsymbol{\xi}_0 = \mathbf{0}$ ,  $\nu = \max(6, d + 4)$ ,  $V = 12I_d$ ,  $\tilde{V} = 2I_d$ ,  $\tilde{\nu} = \nu$ ,  $\tilde{\boldsymbol{\xi}}_0 = \boldsymbol{\xi}_0$ ,  $\tilde{\kappa} = \kappa$  and  $\sigma_{\boldsymbol{\alpha}}^2 = 10$ . Finally, the tempered sequential Monte Carlo algorithm makes use of 10 000 particles.

#### 4.1.1. Parameters estimation

Tables 1 and 2 report respectively the results for the two  $\mathcal{ESN}_1$  distributions (4) and (5) when the sample size is 1 000 and 5 000. Several points are worth commenting. First, as to be expected, the parametrization matters irrespective of the posterior statistics criteria used to compare the overall fitting (posterior mean, posterior median or posterior mode) and of the sample size. More specifically, when the true model is defined from the hidden truncation-based representation (P1), the posterior mean, median and mode using the second parametrization have a larger bias than in the case of the first parametrization. Unsurprisingly, turning to the Bayes factor, we do observe a clear evidence in favor of the results obtained under P1. In contrast, when the true distribution is defined from the convolution-based representation, results in Table 2, and especially the Bayes factor, clearly provide support for the P2-based estimates. This parametric dependence is further illustrated in Figure 1 which displays the marginal posterior distributions using P1 and P2 when the sample size is 1 000.

[Tables 1-2 and Figure 1 here]

Second, comparing Tables 1 and 2, it is worth noting that the results obtained using P1 are less sensitive to the parametrization of the underlying model than those obtained under P2. To understand this point, note that the parameter values of the  $\mathcal{ESN}$  distribution (4) are such that  $\omega^2$  is close to the boundary of the parameter space ( $\omega^2 \approx 0.038$ ) and, consequently, inference for this parameter is very sensitive to the choice of the prior distribution (see e.g. Newton and Raftery, 1994; Gelman, 2006). In particular, the prior we choose for  $\omega^2$  puts a very small weight to values close to zero and therefore tends to overestimate  $\omega^2$ . This nearly boundary problem is critical in the sense that even "non informative" prior distributions can have a substantial effect on inference (see e.g. Gel-



man et al., 2008). For that reason, and contrary to the current practise (see e.g. Adcock, 2004; Liseo and Parisi, 2013), we advocate for the use of the  $(\boldsymbol{\xi}, \Sigma, \boldsymbol{\alpha}, \lambda)$ -parametrization to carry out parameter inference in the ESN (and in the SN) distribution.

However, and this is our third observation, when the sample size gets larger and larger, posterior modes converge toward the true parameter values irrespective of the chosen parametrization. In particular, the middle panel of Figure 1 provides strong support for the convergence of the marginal posterior modes when the sample size is 10 000.

Finally, taking the low number of particles ( $N = 10\,000$ ), the Monte Carlo error is rather small in all cases and for all parameters, especially as the sample size increases. However, it is at the expense of a somehow large computing time which is, for both parametrization, around 90 seconds for  $n = 1\,000$  and around 460 seconds for  $n = 5\,000$ .

#### 4.1.2. Model selection

As explained in Section 3.2.3, it is critical to assess the robustness of ESN distributions with respect to Gaussian distributions. Therefore we conduct some simulation experiments regarding the Bayes factor to test the null hypothesis of normality against the alternative hypothesis of an extended skew-normal distribution,  $\mathcal{ESN}_1^{(P_1)}(2, 6, \boldsymbol{\alpha}, \lambda)$ , for different  $(\boldsymbol{\alpha}, \lambda)$  pairs. The results are reported in Table 3, which describes the percentage of samples where the evidence in favor of the ESN hypothesis is poor ( $\log_{10} B_{10} \leq 0.5$ ), substantial ( $0.5 < \log_{10} B_{10} \leq 1$ ), strong ( $1 < \log_{10} B_{10} \leq 2$ ) and decisive ( $\log_{10} B_{10} > 2$ ).

[Table 3 here]

The results presented in the first three lines of Table 3 are obtained for a sample size of  $n = 100$ . We observe that, despite the small number of observations, the Bayes factor yields very good results for  $(\boldsymbol{\alpha}, \lambda) = (0, -)$  (i.e., Gaussian model) and  $(\boldsymbol{\alpha}, \lambda) = (5, -2)$ . Indeed, in both cases and in all samples, the Bayes factor selects the correct model with a strong confidence. For  $(\boldsymbol{\alpha}, \lambda) = (0.5, 0)$ , estimations are in favour of the Gaussian distribution although the underlying model is ESN. This results is intuitive. Indeed, the Bayes factor penalizes for the number of parameters. Therefore, since  $\lambda$  is useless when the underlying model is Gaussian, it is natural that the Bayes factor is biased toward the Gaussian distribution when  $\alpha$  is close to zero. In contrast, when the sample size increases (from  $n = 100$  to 5 000), the Bayes factor selects the correct model with a

strong confidence. These results suggest that the Bayes factor is convergent even if no formal proof for this specific test is yet available in the literature (see Section 3.2.3).

## 4.2. Example 2: Extended skew-normal sample selection model

One key feature of the proposed methodology is its adaptability since SMC samplers can be used, at least from a theoretical point of view, as soon as one can evaluate efficiently the likelihood function. To illustrate this point in a more complicated set-up than in the previous subsection, we consider the estimation of a sample selection model based on ESN error terms.

### 4.2.1. Model description

Thanks to Definition 1, the application of ESN distributions in sample selection models or Tobit-type models (Amemiya, 1986; Maddala and Lee, 1976) is a natural choice since any hidden truncation of normal component densities leads to such a distribution (see Arnold and Beaver, 2002). In this respect, starting from the Gaussian sample selection model (Heckman, 1976), a (multivariate) extended skew-normal sample selection model (ESNSM) can be defined by:

$$\begin{cases} \mathbf{Y}_i^* = B\mathbf{x}_i + \epsilon_{1i} \\ S_i^* = \beta_2'\mathbf{x}_i + \epsilon_{2i}, \quad i = 1, \dots, n \end{cases} \quad (6)$$

where  $B \in \mathbb{R}^{d \times k_1}$ ,  $\beta_2 \in \mathbb{R}^{k_1}$ , and

$$\epsilon_i \sim \mathcal{ESN}_{d+1}^{(P_1)} \left( \boldsymbol{\xi} = (\xi_1, \xi_2), \Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & 1 \end{pmatrix}, \boldsymbol{\alpha} = (\alpha_1, \alpha_2), \lambda \right) \quad (7)$$

with  $\boldsymbol{\xi} = -\frac{\Sigma\boldsymbol{\alpha}}{c_0} \frac{\phi(\lambda/c_0)}{\Phi(\lambda/c_0)}$  such that  $\mathbb{E}[\epsilon_i] = \mathbf{0}_d$ . We assume that we observe  $S_i = \mathbb{I}_{\mathbb{R}_+}(S_i^*)$  and  $\mathbf{Y}_i = \mathbf{Y}_i^* S_i$ , with  $\mathbb{I}_A(\cdot)$  the indicator function of  $A \subseteq \mathbb{R}$ . The likelihood function of the model, which is required to compute the importance weights of the SMC sampler (Algorithm 1), follows from a direct application of Proposition 1 (closure under conditioning

and marginalization of the extended skew-normal family of distributions):

$$L_n(\theta, \beta_2, B) = \prod_{i=1}^n \left[ \frac{\Phi_2(-\beta_2' \mathbf{x}_i - \xi_2, 1, c_2 \tilde{\alpha}_2, c_2 \lambda)}{\Phi\left(\frac{c_2 \lambda}{\sqrt{1+c_2^2 \tilde{\alpha}_2^2}}\right)} \right]^{1-s_i} \\ \times \left[ \phi_d(\mathbf{y}_i, B\mathbf{x}_i + \xi_1, \Sigma_1) \frac{\Phi_2(m_i, \sigma_{22.2}^2, -\alpha_2, \lambda + \tilde{\alpha}_1'(\mathbf{y}_i - B\mathbf{x}_i - \xi_1))}{\Phi\left(\frac{c_1 \lambda}{\sqrt{1+\tilde{\alpha}_1' \Sigma_1 \tilde{\alpha}_1 c_1^2}}\right)} \right]^{s_i}$$

where  $m_i = \xi_1 + \beta_2' \mathbf{x}_i + \Sigma_{12} \Sigma_1^{-1}(\mathbf{y}_i - B\mathbf{x}_i - \xi_1)$  and with  $\tilde{\alpha}_1$  and  $\tilde{\alpha}_2$  defined as in Proposition 1. The prior distributions for the  $\alpha$  and  $\lambda$  parameters are the same as the ones defined in Section 3.1, while those for  $\Sigma$ ,  $B$  and  $\beta_2$  are discussed in Appendix B.

#### 4.2.2. Simulation set-up

The numerical study is conducted for the univariate extended skew-normal sample selection model:

$$\begin{cases} Y_i^* = \beta_{10} + \beta_{11}x_{1i} + \epsilon_{1i} \\ S_i^* = \beta_{20} + \beta_{22}x_{2i} + \epsilon_{2i}, \end{cases} \quad (8)$$

and

$$(\epsilon_{1i}, \epsilon_{2i}) \sim \mathcal{ESN}_2^{(P_1)}\left(\xi, \begin{pmatrix} 6 & \rho\sqrt{6} \\ \rho\sqrt{6} & 1 \end{pmatrix}, (2, 1), -2\right) \quad (9)$$

where  $\rho \in \{-0.9, 0.3, -0.9\}$ . The parameter value of  $\rho$  is a key issue in sample selection models. Notably when  $\rho = 0$ , there is no selection effect. On the other hand, it can be shown that the correlation between  $\epsilon_{1i}$  and  $\epsilon_{2i}$  increases with this parameter, as shown in Figure 2. The parameter values for the  $\beta$ 's parameters are respectively given by  $\beta_{10} = 3$ ,  $\beta_{11} = -2$ ,  $\beta_{20} = 1.5$  and  $\beta_{22} = 2$  while the covariates  $x_{1i}$  and  $x_{2i}$  are assumed to be independent  $\mathcal{N}_1(0, 2)$  random variables (without loss of generality). This setup implies that  $S_i = 0$  for about 30% – 35% of the  $n = 1\,000$  observations.

[Figure 2 here]

We discuss below some Monte Carlo simulations results for the extended skew-normal sample selection model (8)-(9). The purpose of this numerical study is to compare it with the standard Tobit-type 2 model (i.e., the sample selection model with Gaussian errors, see Amemiya, 1986) regarding the estimation of the parameters and of the marginal effects.

### 4.2.3. Parameters estimation

Table 4 provides the posterior mean and the standard deviation of 50 independent estimates of the parameters of the model (8)-(9) under the two parametric assumptions (i.e., the bivariate extended skew-normal and the Gaussian distribution of the error terms). Results are reported for the three different values of  $\rho$ .

[Table 4 here]

Regarding the estimation of the constant and slope parameters of the regression equation,  $\beta_{10}$  and  $\beta_{11}$ , we observe that the distributional assumption has a limited effect on the estimated values in all scenarios. A similar result is obtained for the Student selection model in Marchenko and Genton (2012) and for the skew-normal model in Ogundimu and Hutton (2012). On the other hand, the estimation of the corresponding parameters in the selection equation,  $\beta_{20}$  and  $\beta_{22}$ , are more sensitive to the choice of the error terms distribution. Indeed, if the Gaussian assumption leads to a small bias for these parameters when the correlation between the variable of interest and the selection variable is low (i.e. when  $\rho = 0.3$ , implying a correlation between -0.02 and -0.03), the results obtained with the Tobit 2 model for these parameters are significantly biased for larger values of  $|\rho|$ .

[Figures 3a-3b here]

To illustrate the importance of the bias for  $\beta_{20}$  and  $\beta_{22}$ , Figure 3a reports, when  $\rho = 0.9$ , the individual estimates of these parameters over 50 simulations in the presence of misspecified error terms—they are wrongly assumed to be normally distributed. Taking that the true parameter vector is given by  $(\beta_{20}, \beta_{22}) = (1.5, 2)$ , we observe that all of the estimates of  $\beta_{20}$  and  $\beta_{22}$  are much larger than the true underlying parameter values. To some extent, this result is consistent with standard results relative to the misspecification issues of the maximum likelihood estimator of Tobit-type models (Amemiya, 1986).

In contrast, when the model is correctly specified, the constant and slope parameters of the selection process are well-estimated irrespective of the correlation parameter  $\rho$ . Notably, the posterior mean of each parameter is close to the true parameter value and the estimation error is small. Regarding other parameters, we obtain very good estimations of  $\sigma_1^2$  and  $\rho$  for which we observe both a small bias and a small standard deviation.

The estimation of  $\alpha_2$  turns out to be more challenging due to the loss of information engendered by the censorship mechanism through  $S_i^*$ . Moreover, the posterior mean of  $\lambda$  is close to the true value at the expense of a relatively large standard deviation (precision), especially with respect to other parameters.

Further evidence is provided by Figure ??, which displays the marginal posterior distributions of the parameters in the case of one realization of the model (8)-(9) with  $\rho = 0.3$ . In addition to previous results, three points are worth commenting. First, the posterior modes are close to the true parameter values. Second, the marginal distribution for the  $\beta$ 's parameters are very concentrated around the mode. Third, the sign of the  $\alpha$ 's parameters, and hence of the skewness of the data, is well-identified since the posterior mass on  $\{\alpha_i < 0, i = 1, 2\}$  is close to zero. In contrast, there is a small but significant posterior probability for the event  $\{\lambda > 0\}$  suggesting that more observations are needed to identify more precisely this parameter.

[Figures 4 here]

#### 4.2.4. Marginal effects

For ease of interpretation, it is arguably better to consider the (average) marginal effects (Cameron and Trivedi, 2005) since only the sign (but not the magnitude) of the coefficients can be readily interpreted in Tobit-type models. In this respect, we compute the marginal effects (see Proposition 3 in Appendix B) and Figure 4 displays marginal effect estimates of  $\beta_{22}$  on  $\mathbb{E}[Y_i^* | S_i = 1, \mathbf{x}_i]$  for a realization of the above model with  $\rho = 0.3$ . The main result is that the Gaussian model is not able to account for substantial heterogeneity in marginal effects. Indeed, a visual inspection shows that the distribution of Gaussian estimates is much more concentrated than the distribution of the true values. In addition, the marginal effects obtained from the Tobit type-2 models are in all cases larger than -10 although for a very significant proportion of individuals the marginal effect of  $\beta_{22}$  is indeed smaller than this threshold (with a minimum nearby -60). The average marginal effect estimate under the Gaussian assumption is around -0.14 while the true value is about 15 times larger (around -2.08). In contrast, this estimate under the ESN assumption is -2.22. Some marginal effect estimates under the correct parametric assumption are also reported and are, as to be expected, very close to the true values.

## 5. Applications

In this section, we proceed with two real applications. In both cases, estimations are performed using the P1-parametrization (Definition 1), with the prior distribution as in Section 4.1. The absence of a constraint on  $\lambda$  contrasts with most of the applications of skew-elliptical distributions in the literature that set the  $\lambda$  parameter to zero (or consider some arbitrary value of  $c$ ). Finally, the SMC sampler is initialized using a pilot run of a Gaussian random walk Metropolis-Hastings and the propagation step is based on  $\tau = 3$  iterations of a Gaussian Metropolis-Hastings kernel (see Section 3.2).

### 5.1. Transfer fees of soccer players

As an application of the univariate ESN, we consider a data set with 1 062 observations on (log-) transfer fees in major European soccer leagues. Data, which have been collected from various sources and are available upon request, cover the period 2008-2012 for the first league (England and France), Bundesliga (Germany), Calcio (Italy) and Liga (Spain).

[Figures 5 and 6 here]

Figure 5 (resp. Figure 6) presents the marginal posterior distributions when data are assumed to be randomly generated from an ESN distribution (resp. a SN distribution). Visual inspection of the marginal posterior distribution indicates that the ESN-based marginal posterior distribution of  $\lambda$  has most of its mass on the interval  $(-\infty, -2]$ . Paired with the fact that the posterior distribution for  $\alpha$  has most of its mass on  $[1.2, \infty)$ , this suggests that the ESN-based specification fits better the overall distribution of data than the SN-based specification of Azzalini (1985). Further evidence can then be provided by comparing the marginal likelihood values of the two models. Notably, using the output of the SMC sampler, the (log-) evidence of the ESN-based model is -685.0374 whereas it is only -797.1437 in the case of the SN-based model. This means that the evidence in favor the ESN-based model can be considered as being "decisive" in the sense of Jeffreys (1939). Finally, to assess the robustness of our results, Figure 7 compares the ESN estimate of the density function of the data with a non-parametric estimate: one can observe that both provide very similar results.

[Figure 7 here]

## 5.2. Bivariate ESN: Financial Data

As a final illustration of the proposed algorithm, we proceed with a real financial data set as in Liseo and Parisi (2013). There is an impressive literature in finance that has witnessed the fact that (high-frequency) financial returns are skewed and display leptokurtic tails (e.g., see Jondeau et al., 2006; Genton, 2004) and may have strong implications in portfolio selection, asset pricing models or risk measurement (among others). In this respect, we consider a simple i.i.d. bivariate sampling model. More specifically, we analyse the weekly returns (in percentage) of two US stocks, namely “ABM Industries Incorporated” (ABM) and “The Boeing Company” (BA). The sample size covers the period Jul 19, 1984 to Jul 28, 2014 (1 566 observations).<sup>1</sup>

[Figures 8 and 9 here]

Figures 8 and 9 depict, respectively, the marginal posterior distribution of each parameter under the ESN and the SN assumption. Two points are worth commenting. First the contour plot of the density of the estimated  $\mathcal{ESN}_2$  model suggests that raw data, which are skewed and fat-tailed, can be reasonably well-captured by this specification. Second, the marginal posterior modes of the shape ( $\alpha_1$  and  $\alpha_2$ ) and the shift parameter ( $\lambda$ ) are roughly given by 0.13, 0.20 and -3, respectively. Combined with the fact that the (marginal) posterior of each of these parameters has a negligible mass with positive (for  $\lambda$ ) or negative (for  $\alpha_1, \alpha_2$ ) values, the estimation provides strong support for the application of an extended skew-normal distribution in order to jointly model ABM and BA. Moreover, according to standard stylized financial facts of weekly returns, the location parameters,  $\xi_1$  and  $\xi_2$ , are negative (large negative returns are more important than large positive returns) and the marginal posterior modes of the unconditional variance-covariance parameters ( $\sigma_1^2, \sigma_2^2, \sigma_{12}^2$ ) support large volatility and co-volatility.

Finally we proceed with model selection. Using the SMC estimate of the Bayes factor, we find that the evidence in favor of the skew-normal bivariate distribution proposed by Liseo and Parisi (2013) is poor (in the sense of Jeffreys, 1939).

---

<sup>1</sup>We also perform estimation with daily and monthly returns. Our main results remain unchanged.

## 6. Conclusion

In this paper, we propose a new Bayesian computational approach, which rests on a tempered sequential Monte Carlo sampler, to estimate (multivariate) extended skew-normal distributions. Among others, the proposed approach have several advantages. First, it overcomes some issues encountered in standard maximum likelihood estimation. Second, in contrast to MCMC methods, it is easy to build a SMC algorithm that is adaptive in the sense that it can adjust sequentially and automatically its sampling distribution to the problem at hand provided some well-defined prior distributions. Especially, it can implemented for a large class of (multivariate) skew-elliptical distributions. Third, it allows to compute easily as a by-product the marginal posterior distributions, the normalizing constant and thus the Bayes factor. Fourth, it embeds as a special case the population algorithm provided by Liseo and Parisi (2013).

Monte Carlo simulations provide evidence regarding the robustness of the proposed algorithm with different data generating processes. Irrespective of the model considered (sampling models, extended skew-normal sample selection models), posterior statistics are rather precise (with a low standard deviation) in a tractable computing time. Moreover, results suggest that the hidden truncation-based parametrization is more robust for estimation than the convolution-based parametrization. Directions for future research include more comprehensive empirical applications (Gerber and Pelgrin, 2014) and the derivation of more general models with hidden truncation, censoring or selective report with the (multivariate) extended skew-normal family of distributions or some unified skew-elliptical distributions.

## Acknowledgements

Financial support from the Swiss National Science Foundation (research module “Modelling simultaneous equation models with skewed and heavy-tailed distributions”) is gratefully acknowledged.



## A. Proof of Proposition 2

Let  $l_n(\theta) = \log L_n(\theta)$  and  $\theta \in \Theta_\epsilon^{l^*}$  where  $\Theta_\epsilon(l^*) = \{\theta : \|\theta - \theta_{n,G}^{l^*}\| \leq \epsilon\}$ . Then,  $l_n(\theta_{n,G}^{l^*}) - l_n(\theta) > 0$  means that

$$l_n^G(\theta_{n,G}^{l^*}) - l_n^G(\theta) - \frac{1}{N} \sum_{i=1}^n \log \Phi(l + a(z_i - m)) + \log \Phi(l/c_0) \geq 0$$

where  $l_n^G$  is the log-likelihood corresponding to the Gaussian model. A sufficient condition for the above inequality to hold is

$$\log \Phi \left( \frac{l^* - \epsilon}{\sqrt{1 + (\sigma_{n,G}^2 + \epsilon)\epsilon^2}} \right) \geq \log \Phi(l^* + \epsilon(1 + \bar{z}_n - \xi_{n,G} + \epsilon))$$

where  $\bar{z}_n = \max\{|z_i|\}$ . This is equivalent to

$$l^* \leq l_{n,\epsilon}^* := \frac{\epsilon + \sqrt{1^* + (\sigma_{n,G}^2 + \epsilon)\epsilon^2} (\epsilon(1 + \bar{z}_n - \xi_{n,G} + \epsilon))}{1 - \sqrt{1^* + (\sigma_{n,G}^2 + \epsilon)\epsilon^2}}.$$

Hence, for all  $\epsilon > 0$ , there exists a  $l_{n,\epsilon}^*$  such that

$$l_n^G(\theta_{n,G}^{l_{n,\epsilon}^*}) - l_n^G(\theta) - \frac{1}{N} \sum_{i=1}^n \log \Phi(l + a(z_i - m)) + \log \Phi(l/c_0) \geq 0 \quad \forall \theta \in \Theta_\epsilon(l_{n,\epsilon}^*).$$

To prove part 2., let  $\epsilon$  and  $M \geq 1$  be such that  $c_n := \|\tilde{\theta}_n - \tilde{\theta}_G\| = \frac{\epsilon}{M}$  where  $\theta = (\tilde{\theta}, l)$ . Then, if

$$l_{n,\epsilon}^* - \epsilon \left[ 1 - \frac{c_n^2}{\epsilon^2} \right]^{1/2} \leq l_n \leq l_{n,\epsilon}^* + \epsilon \left[ 1 - \frac{c_n^2}{\epsilon^2} \right]^{1/2}$$

we have  $\|\theta_n - \theta_G^{l_{n,\epsilon}^*}\| \leq \epsilon$  so that  $l_n(\theta_G^{l_{n,\epsilon}^*}) \geq l_n(\theta_n)$ .

## B. Extended skew-normal sample selection models

### B.1. Prior distributions for $\Sigma$ , $B$ and $\beta_2$

When available, the conjugate prior distribution is frequently used in bayesian analysis. Under Gaussian error terms and no selection effect, the conjugate prior distribution for  $\beta_1$  and  $\Sigma$  is the normal-inverse Wishart distribution:

$$\pi(\beta_1, \Sigma | \mu_{\beta_1}, \kappa, \nu, V) \propto \exp \left( -\frac{1}{2} \text{tr}(V \Sigma^{-1}) - \frac{\kappa}{2} (\beta_1 - \mu_{\beta_1})' (\Sigma^{-1} \otimes c_{\beta_1} X' X) (\beta_1 - \mu_{\beta_1}) \right)$$

$$\times |\Sigma|^{-\frac{\nu+|\beta_1|+2}{2}}$$

where  $c_{\beta_1}$  is a scale factor,  $V$  is a  $d \times d$  positive definite matrix,  $\kappa$  and  $\nu$  are real such that  $\nu > |\beta_1| + 3^2$ . Since the ESN distribution generalizes the Gaussian distribution, and because the presence of selection effect does not modify our prior knowledge, we choose this prior distribution for  $\beta_1$  and  $\Sigma$ .

Using a similar argument, a possible choice of prior distribution for the parameters of the selection equation is  $\beta_2 \sim \mathcal{N}_{|\beta_2|}(\mu_{\beta_2}, c_{\beta_2}(X'X)^{-1})$  where  $c_{\beta_2}$  is a scale factor. This choice of prior distribution for  $(\beta, \Sigma)$  is particularly convenient for model selection because under Gaussian error terms and no selection effect the posterior mean of  $\beta_1$  (respectively,  $\Sigma$ ) has a closed form expression provided that  $\beta_1$  and  $\beta_2$  are *a priori* independent. In the numerical study (Section 4.2), parameters of prior distributions are given by  $\mu_{\beta_1} = \mu_{\beta_2} = 0$  and  $c_{\beta_1} = c_{\beta_2} = 5n$ , with  $n$  the sample size.

## B.2. Determination of the marginal effects

**Proposition 3.** *Consider the univariate extended skew-normal sample selection model defined by (6) and (7). Let*

$$\tau(a, \alpha, \lambda) = \frac{\phi(a)\Phi(\lambda + \alpha a)}{\Phi_2(a, 1, \alpha, \lambda)}, \quad \delta(a, \alpha, \lambda) = \frac{\phi(\lambda/c_0)\Phi\left(ac_0 + \frac{\alpha\lambda}{c_0}\right)}{\Phi_2(a, 1, \alpha, \lambda)}.$$

Then,

$$\begin{aligned} \mathbb{E}[S_i^* | S_i = 1, \mathbf{x}_i] &= \beta_2' \mathbf{x}_i + \tau_{2i} + \frac{c_2 \tilde{\alpha}_2}{c_{02}} \delta_{2i} \\ \mathbb{E}[Y_i^* | S_i = 1, \mathbf{x}_i] &= \xi_{1i} + \mathbf{x}_i' \beta_1 + \sigma_{12} \tau_{2i} + \sigma_1 v_2 \delta_{2i} \end{aligned}$$

where  $\tau_{2i} = \tau(\xi_2 + \mathbf{x}_i' \beta_2, -c_2 \tilde{\alpha}_2, c_2 \lambda)$ ,  $\delta_{2i} = \delta(\xi_2 + \mathbf{x}_i' \beta_2, -c_2 \tilde{\alpha}_2, c_2 \lambda)$  and  $v_2 = \frac{\rho c_2 \tilde{\alpha}_2 + c_2 (1 - \rho^2) \alpha_1}{c_{02}}$ .

Proof: See Gerber and Pelgrin (2014).

---

<sup>2</sup>This last condition is not necessary but ensures that all the components of  $\Sigma$  has a finite variance.

## C. Tables

Table 1: Estimation of univariate  $\mathcal{ESN}_1$  distributions (4)

	Estimation under P1			Estimation under P2		
	Mean	Median	Mode	Mean	Median	Mode
$\xi = 2$	-15%	-13%	-11.5%	-73.5%	-73%	-72.5%
	(0.010)	(0.009)	(0.012)	(0.008)	(0.009)	(0.012)
	-8%	-7.5%	-7%	-31%	-30.5%	-30%
	(0.002)	(0.003)	(0.005)	(0.003)	(0.004)	(0.006)
$\sigma^2 = 6$	-2.3%	-3.5%	-4.3%	25.3%	24.7%	24.2%
	(0.017)	(0.016)	(0.019)	(0.011)	(0.013)	(0.016)
	3.7%	3.4%	3.2%	15.5%	15.3%	15.2%
	(0.004)	(0.005)	(0.009)	(0.005)	(0.006)	(0.010)
$\alpha = 5$	-19.8%	-20.2%	-20.4%	-39.2%	-39.8%	-40.0%
	(0.006)	(0.006)	(0.010)	(0.006)	(0.005)	(0.007)
	-5.8%	-6.0%	-6.2%	-19.4%	-19.6%	-19.6%
	(0.003)	(0.004)	(0.006)	(0.002)	(0.003)	(0.005)
$\lambda = -2$	55%	47%	42%	212%	207.5%	204%
	(0.040)	(0.036)	(0.041)	(0.023)	(0.024)	(0.029)
	38%	35.5%	34%	118%	116.5%	115%
	(0.011)	(0.014)	(0.021)	(0.011)	(0.013)	(0.022)
$\log m(z_{1:n})$	-1 473.84			-1 532.98		
	(38.37)			(34.00)		
	-8 065.57			-8 074.10		
Time (in seconds)	(13.01)			(18.22)		
	60.22			34.05		
	120.44			124.26		

Notes: The results are obtained from 50 estimations of the model with  $N = 10\,000$  particles. Mean estimates are reported as percentage deviation of the true parameter value, and standard deviations are given in brackets. For each parameter, the first (respectively, last) two rows correspond to  $n = 1\,000$  (respectively,  $n = 5\,000$ ).

Table 2: Estimation of univariate  $\mathcal{ESN}_1$  distributions (5)

	Estimation under P1			Estimation under P2		
	Mean	Median	Mode	Mean	Median	Mode
$\xi = 2$	76%	83%	88.5%	-10.5%	-9%	-7%
	(0.031)	(0.030)	(0.035)	(0.017)	(0.019)	(0.029)
	50.5%	53.5%	55.5%	26%	27%	28%
	(0.010)	(0.012)	(0.022)	(0.010)	(0.010)	(0.014)
$\sigma^2 = 26$	-12.7%	-14.1%	-15.2%	6.8%	6%	5.4%
	(0.092)	(0.082)	(0.094)	(0.045)	(0.051)	(0.071)
	-4.4%	-5%	-5.5%	1.1%	0.9%	0.6%
	(0.029)	(0.037)	(0.055)	(0.024)	(0.030)	(0.046)
$\alpha = 0.98$	2%	2%	2%	2%	-2%	0%
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
	3.1%	4.1%	4.1%	4.1%	4.1%	4.1%
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
$\lambda = -4.08$	-40.2%	-43.1%	45.8%	6.1%	4.4%	3.2%
	(0.032)	(0.031)	(0.034)	(0.019)	(0.019)	(0.033)
	-25.2%	-26.5%	-27.5%	-12.5%	-13%	-13.2%
	(0.010)	(0.013)	(0.021)	(0.009)	(0.010)	(0.018)
$\log m(z_{1:n})$	-2 244.49			-2 184.73		
	(41.69)			(30.01)		
	-11 500.13			-11 476.46		
Time (in seconds)	(25.22)			(41.65)		
	38.52			61.40		
	125.98			165.89		

Notes: The results are obtained from 50 estimations of the model with  $N = 10\,000$  particles. Mean estimates are reported as percentage deviation of the true parameter value, and standard deviations are given in brackets. For each parameter, the first (respectively, last) two rows correspond to  $n = 1\,000$  (respectively,  $n = 5\,000$ ).

Table 3: Bayes factors

	$(\alpha, \lambda)$	$\log_{10} B_{10} \leq 0.5$	$0.5 < \log_{10} B_{10} \leq 1$	$1 < \log_{10} B_{10} \leq 2$	$\log_{10} B_{10} > 2$
n=100	(0,-)	100%	0%	0%	0%
	(5,-2)	1%	1%	4%	96%
	(0.5,1)	100%	0%	0%	0%
n=5 000	(0.5,1)	0%	0%	0%	100%

Notes: The results are obtained from 100 samples. The number of particles is 10 000 and  $B_{10}$  denotes the Bayes factor to test the normality hypothesis.

Table 4: Estimation of sample selection model (8)-(9)

Parameter	$\rho$	Tobit 2		ESNM		True value
		Mean	Standard deviation	Mean	Standard deviation	
$\beta_{10}$	0.3	2.92	0.0008	2.94	0.0006	3
	0.9	2.98	0.0006	2.99	0.0005	
	-0.9	2.97	0.0005	2.98	0.0006	
$\beta_{11}$	0.3	-1.98	0.0004	-1.96	0.0004	-2
	0.9	-1.99	0.0004	-1.99	0.0003	
	-0.9	-1.99	0.0003	-1.990	0.0352	
$\beta_{20}$	0.3	1.58	0.0010	1.37	0.0015	1.5
	0.9	2.57	0.0020	1.78	0.0021	
	-0.9	2.10	0.0015	1.43	0.0020	
$\beta_{22}$	0.3	2.04	0.0013	1.77	0.0020	2
	0.9	3.32	0.0026	2.30	0.0026	
	-0.9	2.77	0.0020	1.87	0.0028	
$\sigma_1^2$	0.3	2.22	0.0012	6.04	0.0101	(6)
	0.9	2.10	0.0011	5.74	0.0074	
	-0.9	1.60	0.0008	6.08	0.0113	
$\rho$	0.3	0.06	0.0010	0.39	0.0015	
	0.9	0.63	0.0010	0.82	0.0006	
	-0.9	-0.76	0.0008	-0.90	0.0004	
$\alpha_1$	0.3	-	-	3.04	0.0154	(2)
	0.9	-	-	3.27	0.0165	
	-0.9	-	-	1.86	0.0066	
$\alpha_2$	0.3	-	-	2.17	0.0145	(1)
	0.9	-	-	2.15	0.0251	
	-0.9	-	-	0.55	0.0134	
$\lambda$	0.3	-	-	-2.54	0.0234	(-2)
	0.9	-	-	-1.85	0.0207	
	-0.9	-	-	-3.38	0.0164	
$\log m(z_{1:n})$	0.3	-1 448.91	0.00251	-1 313.91	0.0269	
	0.9	-1 358.97	0.00401	-1 206.92	0.0504	
	-0.9	-1 291.19	0.0059	-1 168.48	0.0262	

Note: Using  $N = 10\,000$  particles, results are obtained from independent 50 independent estimations.



## D. Figures

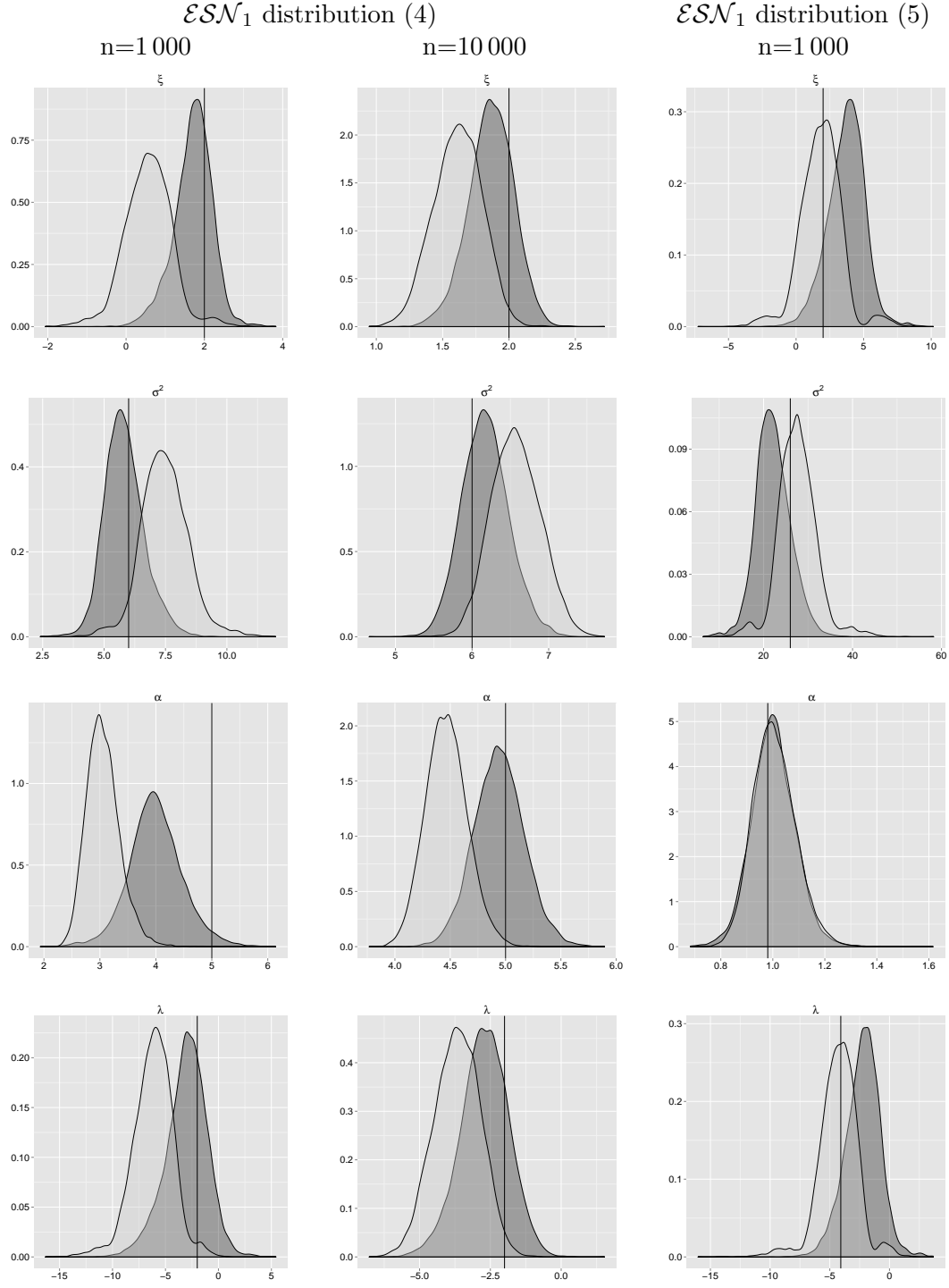


Figure 1: Marginal posterior distributions for the parameters of the  $\mathcal{ESN}_1$  distributions (4) and (5). The results for P1 (respectively, P2) are in dark (respectively, in grey) and are obtained with  $N = 10\,000$  particles.

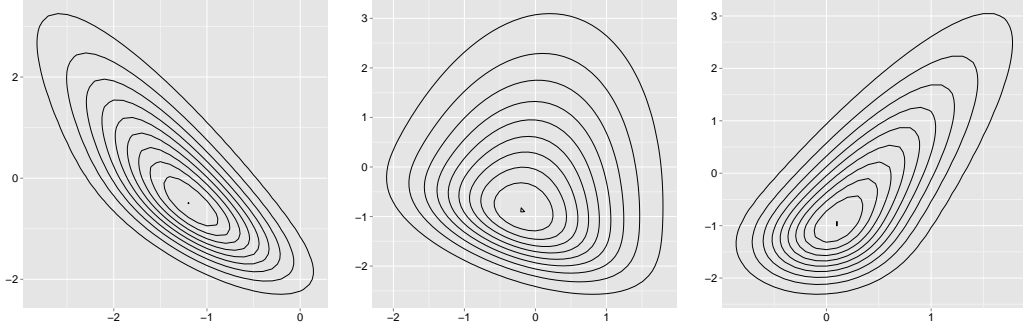
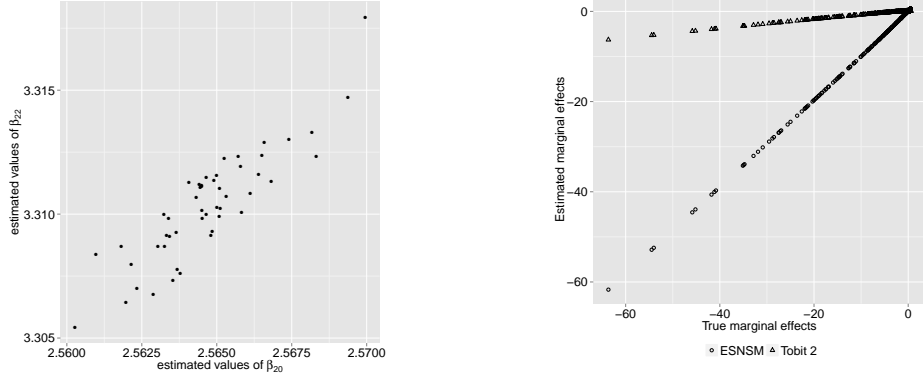


Figure 2: Contour plots of the zero mean  $\mathcal{ESN}_2$  distribution (9). The left, middle and right panels represent the contour plots of the zero mean  $\mathcal{ESN}_2$  distribution when the correlation parameter  $\rho$  is respectively given by -0.9, 0.3, and 0.9.



- (a) The results are obtained using  $N = 10\,000$  particles, 50 independent estimates of the selection parameters  $(\beta_{20}, \beta_{22})$  are reported when the true parameter vector is  $(\beta_{20}, \beta_{22}) = (1.5, 2)$  and  $\rho = 0.9$ .
- (b) The results are obtained with  $N = 50\,000$  particle and  $\rho = 0.3$ . For each individual, we report the marginal effect estimate using either the Tobit type-2 model (triangular markers) or the ESNSM model (circular markers) against the true marginal effect.

Figure 3: Bias for selection coefficients of a Tobit type-2 model (Figure 3a) and marginal effects for the ESNSM (8)-(9) (Figure 4).



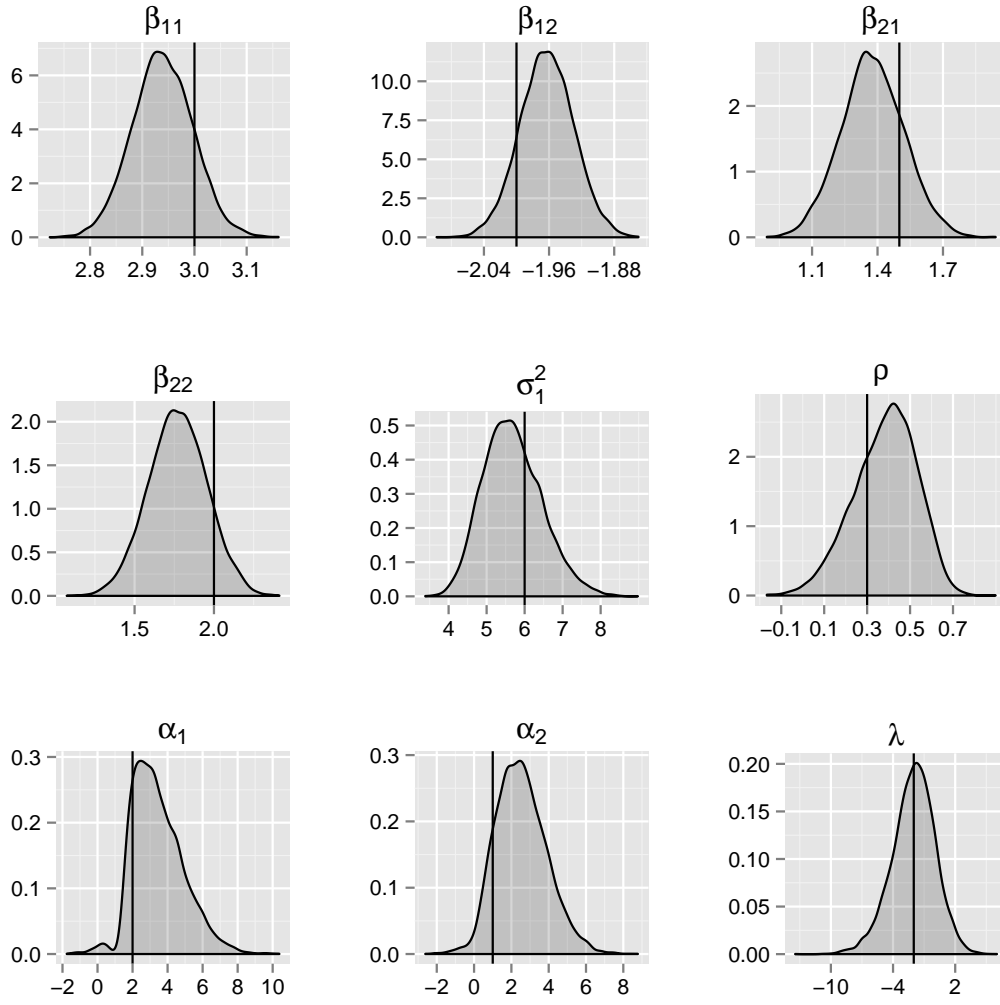


Figure 4: Marginal Posterior distribution of the ESNSM (8)-(9), evaluated with 50 000 particles when  $\rho = 0.3$ .

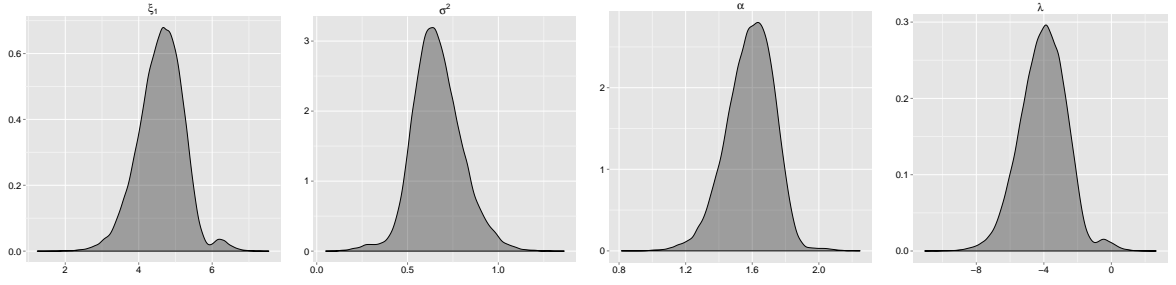


Figure 5: Transfer fees of soccer players and marginal posterior distributions of the SN parameters. The results are obtained with  $N = 50\,000$  particles.

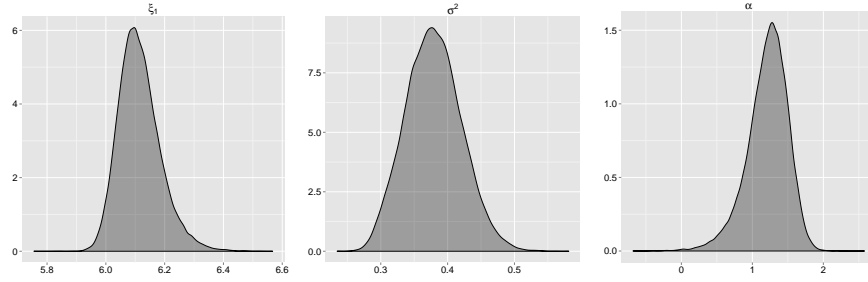


Figure 6: Transfer fees of soccer players and marginal posterior distributions of the SN parameters. The results are obtained with  $N = 50\,000$  particles.

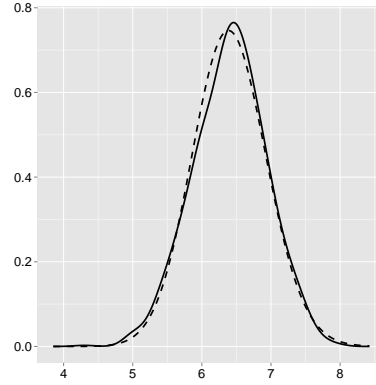


Figure 7: Estimates of the Non-parametric and ESN transfer fees distribution. The dashed line (respectively, solid line) represents the estimate of the ESN (respectively, non-parametric) transfer fees distribution. The ESN estimate is obtained with  $N = 50\,000$  particles.

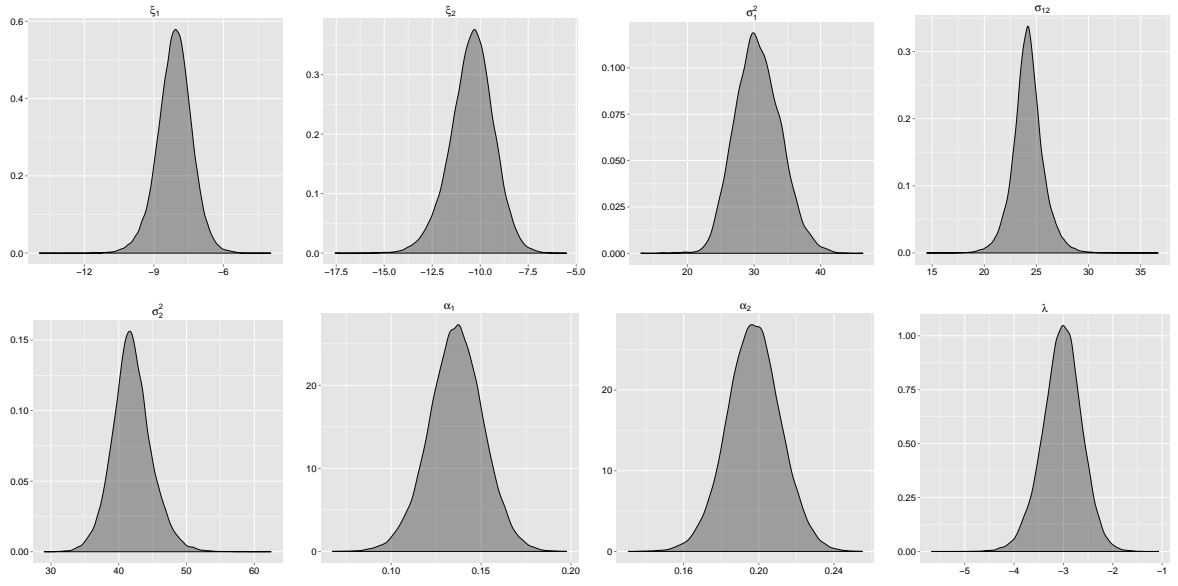


Figure 8: Marginal Posterior distributions for the financial data under the ESN assumption. The results are obtained with  $N = 10\,000$  particles. Evidence (in log) is -8 631.379.

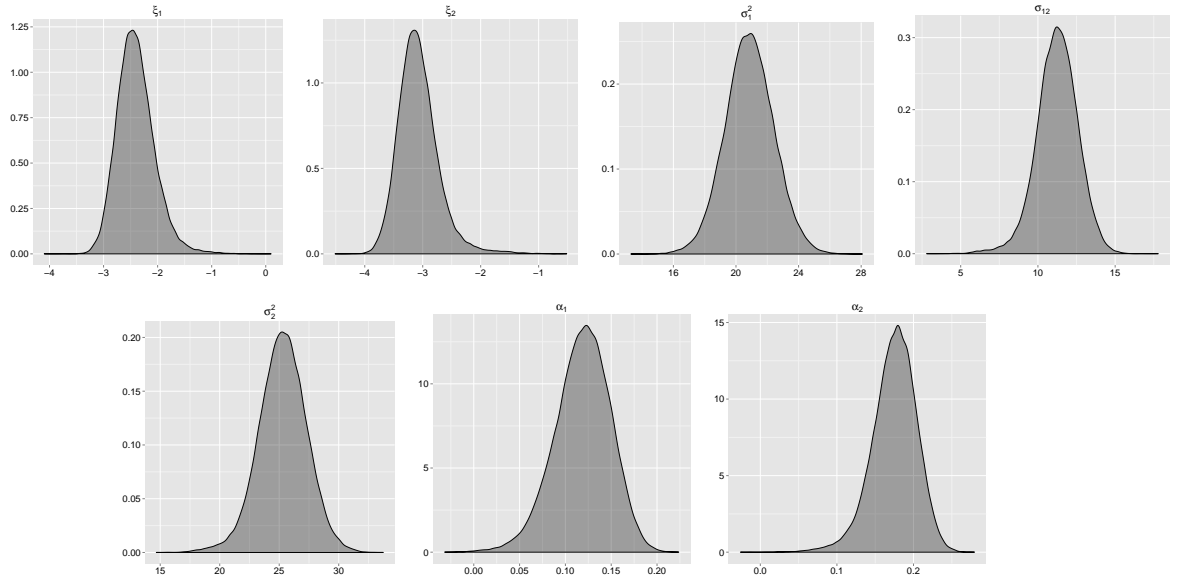


Figure 9: Marginal Posterior distributions for the financial data under the SN assumption. The results are obtained for  $N = 50\,000$ . Evidence (in log) is -8 647.239.

## References

- Adock, C. (2004). Capital asset pricing for UK stocks under the multivariate skew-normal distribution. In Genton, M. G., editor, *Skew-elliptical distributions and their applications*, pages 191–204. Chapman & Hall/CRC.
- Amemiya, T. (1986). *Advanced Econometrics*. Princeton University Press.
- Arellano-Valle, R. B., Branco, M. D., and Genton, M. G. (2006). A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics*, 34(4):581–601.
- Arellano-Valle, R. B. and Genton, M. G. (2010). Multivariate unified skew-elliptical distributions. *Chilean Journal of Statistics*, 1(1):17–33.
- Arnold, B. C. and Beaver, R. J. (2000). The skew-cauchy distribution. *Statistics and Probability Letters*, 49(3):285–290.
- Arnold, B. C. and Beaver, R. J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Mathematics and Statistics*, 11(1):7–54.
- Arnold, B. C., Beaver, R. J., Groeneveld, R. A., and Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, 58(3):471–478.
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178.
- Azzalini, A. and Capitanio, A. (1998). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602.
- Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.
- Branco, M. D., Genton, M. G., and Liseo, B. (2013). Objective Bayesian analysis of skew-t distributions. *Scandinavian Journal of Statistics*, 40(1):63–85.

- Cabral, C., Lachos, V., and Prates, M. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics and Data Analysis*, 56(1):126–142.
- Cameron, C. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.
- Capitanio, A., A. A. and Stanghellini, E. (2003). Graphical models for skew-normal variates. *Scandinavian Journal of Statistics*, 30:129–144.
- Carpenter, J., Clifford, P., and Fearnhead, P. (1999). An improved particle filter for non-linear problems. *IEEE Proceedings - Radar, Sonar and Navigation*, 146(1):2–7.
- Celeux, G., Marin, J.-M., and Robert, C. P. (2004). Iterated importance sampling in missing data problems. *Computational statistics & data analysis*, 50(12):3386–3404.
- Copas, J. and Li, H. (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:55–95.
- Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.
- Domínguez-Molina, J., González-Farías, G., and Gupta, A. K. (2003). The multivariate closed skew normal distribution. Technical Report 03-12, Department of Mathematics and Statistics, Bowling Green State University.
- Fang, B. Q. (2003). The skew elliptical distributions and their quadratic forms. *Journal of Multivariate Analysis*, 87(2):298–314.
- Früwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2):317–336.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian analysis*, 1(3):515–533.
- Gelman, A., Carlin, J., Stern, H., and Rubin, M. (2004). *Bayesian data analysis*. Chapman and Hall.

- Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.
- Gelman, A. and Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13.
- Genton, M. G., editor (2004). *Skew-elliptical distributions and their applications*. Chapman & Hall/CRC.
- Gerber, M. and Pelgrin, F. (2014). The class of skewed-elliptical sample selection models. *Mimeo*.
- Heckman, J. (1976). The common structure of statistical models of truncation sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492.
- Huguenin, J., Pelgrin, F., and Holly, A. (2014). Estimation of multivariate probit models by using a general decomposition of multivariate normal probabilities. Technical report.
- Jeffreys, H. (1939). *Theory of Probability*. Oxford University press, London, third edition.
- Jondeau, E., Poon, S., and Rockinger, M. (2006). *Financial modeling under non-normality*. Springer Verlag, first edition.
- Liseo, B. and Loperfido, N. M. R. (2006). A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical Planning and Inference*, 136(2):373–389.
- Liseo, B. and Parisi, A. (2013). Bayesian inference for the multivariate skew-normal model: a Population Monte Carlo approach. *Computational Statistics and Data Analysis*, 63:125–138.
- Liu, J. and Chen, R. (1995). Sequential monte carlo for dynamic systems. *Journal of American Statistical Association*, 93.
- Maddala, G. S. and Lee, L. F. (1976). Recursive models with quantitative endogenous variables. *Annals of Economic and Social Measurement*, 5(4):525–545.

- Manning, W., Basu, A., and Lee, L. F. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, 24:465–488.
- Marchenko, Y. V. and Genton, M. G. (2012). A heckman selection-t model. *Journal of the American Statistical Association*, 107(497):304–317.
- Morin, J., Pillai, N., Robert, C., and Rousseau, J. (2013). Relevant statistics for bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Neal, R. (2001). Annealed importance sampling. *Statist. Comput.*, 11.
- Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48.
- Ogundimu, E. O. and Hutton, J. L. (2012). A general sample selection model with skew-normal distribution. Technical Report 12-05, Centre for statistical Methodology Warwick.
- O’Hagan, A. and Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63:201–212.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, chapter 9. Springer.
- Sahu, S. K., Dey, D. K., and D’Elia Branco, M. (2003). A new class of multivariate skew distributions with applications to bayesian regression models. *The Canadian journal of Statistics*, 31(2):129–150.
- Schäfer, C. and Chopin, N. (2013). Adaptive monte carlo on binary sampling spaces. *Statistics and Computing*, 23(2):163–184.
- Wiper, M., Gíron, F., and Pewsey, A. (2008). Objective bayesian inference for the half-normal and half-t distributions. *Communications in Statistics: Theory and Methods*, 37(18-20):3165–3185.